

An Alternate Method for Organizing Biological Functions using Gene Annotation Networks

Kimberly Glass^{1,2,3,*}, and Michelle Girvan^{3,4,5}

¹*Department of Biostatistics,*

Harvard School of Public Health, Boston, MA, USA

²*Department of Biostatistics and Computational Biology,*

Dana-Farber Cancer Institute, Boston, MA, USA

³*Department of Physics, University of Maryland,
College Park, MD, USA*

⁴*Institute for Physical Science and Technology,
University of Maryland, College Park, MD, USA*

⁵*Santa Fe Institute, Santa Fe, NM*

The Gene Ontology (GO) provides biologists with a controlled terminology that describes how genes are associated with functions and how functional terms are related to each other. These term-term relationships encode how scientists conceive the organization of biological functions, and they take the form of a directed acyclic graph (DAG). Here, we propose that the network structure of gene-term annotations made using GO can be employed to establish an alternate natural way to group the functional terms which is different from the hierarchical structure established in the GO DAG. Instead of relying on an externally defined organization for biological functions, our method connects biological functions together if they are performed by the same genes, as indicated in a compendium of gene annotation data from numerous different biological experiments. Grouping terms by this alternate scheme provides a new framework with which to describe and predict the functions of experimentally identified sets of genes.

1. INTRODUCTION

The Gene Ontology (GO) [4][29] has been around for over a decade, during which time it has been widely utilized both to validate and to predict the results of biological experiments (see, for example [9, 14, 16, 17, 21, 33, 34]). The structure of the ontology, where different “categories” or terms are related to each other in a hierarchical fashion, provides a well-established format with which to classify and subclassify all biological functions and processes. This classification approach is well-structured and well-characterized, however, we seek to determine if it is the only natural way in which to classify this type of biological information, or if any other, alternate natural groupings of functions might exist. We address two main questions. First, does there exist another natural way to organize the functional terms that is distinct from the ontological organization? Secondly, if such an alternate classification exists, can it be used to interpret biological data?

Our approach is outlined in Figure 1. We begin by considering the term relationships as defined by the GO hierarchy. We then add in annotation information that reflects gene-term relationships obtained from numerous different biological experiments. We encapsulate these connections in the form of a bipartite network. Next, we use this bipartite network to construct another network describing the relationships between the functional terms based on shared gene annotations. We then apply

standard community structure finding algorithms to partition this annotation-driven network into communities of terms. We compare these communities to the previously defined branches from the GO hierarchy and show that, although there are some similarities, there are also very strong differences between the two ways of organizing functional terms. Finally, using functional enrichment techniques, we find a strong association between both our functional term communities and GO branches (ontological groupings of terms) with experimentally-derived sets of genes.

In the past there has only been minimal investigation of how biological functions might be related to each other outside of the established ontology structure, and the majority of this has focused on discovering individual links between functions [16] rather than investigating the structure as a whole. In this work, we propose an alternate, viable classification of functional terms that relies on the network structure of gene-term annotations rather than ontological relationships. We test the applicability of this classification, utilizing functional analysis techniques in order to evaluate the enrichment of cancer signatures (sets of genes associated with a particular cancer) in both term communities and GO branches. We find that certain signatures are highly enriched in the found communities and not GO branches. We therefore suggest that by linking GO functional terms based on shared genes, we can create an alternate, biologically meaningful, network-derived organization of the functional terms that is both distinct from the established GO DAG and can also be used to investigate biological systems.

*contact: kglass@jimmy.harvard.edu

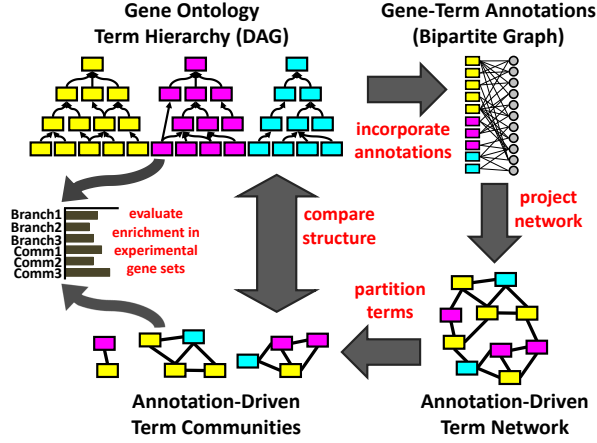


FIG. 1: Visual representation of our approach. First, we summarize gene annotations made to functional terms in the Gene Ontology Hierarchy as a gene-term bipartite graph. From these gene-term relationships, we create a projected term-term network. We partition the term-term network into communities and compare those term communities to branches of terms in the DAG. Finally, we perform functional enrichment analysis on gene sets using both these term communities and GO branches.

2. BACKGROUND

2.1. Identifying and Comparing Community Structure in Networks

In recent years, complex networks tools have been used alongside traditional bioinformatics techniques to study many different kinds of biological networks [24], including, but not limited to, gene regulatory networks [20, 28], protein-protein interaction networks [15, 31], and metabolic networks [13, 35]. Developments in network theory provide the computational tools needed to calculate global properties of such networks, lending insights into the behavior of the systems represented by these networks. Many networks exhibit community structure, meaning that there are clusters of nodes in the network within which there are many edges but between which there are few edges. Recent developments in network theory have been able to detect such functional modules [23] in a computationally efficient and accurate manner [6]. Algorithms for detecting network structure, although most fully developed for un-weighted, un-directed graphs, can also be applied to weighted [22] or directed [18] networks with only slight modifications.

In order to quantify the strength of community structure we use a quantity known as modularity [22, 23]. Modularity (Q) can be defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \left(1 + \frac{r}{\langle k \rangle} \right) \frac{k_i k_j}{2m} \right] \delta(x_i, x_j) \quad (1)$$

where δ is the Kronecker delta function, x_i is the com-

munity of node i , k_i is the degree of node i , A is the adjacency matrix, a matrix with values representing the weight between nodes i and j , and m is the total weight of the edges in the network [3]. Traditionally, in order to partition the network into communities, the resolution parameter, r in Equation 1, is set equal to zero. Varying this value allows one to look for alternate divisions of a network into communities at different scales, with $r > 0$ uncovering sub-structures in the network, finding new communities, or breaking apart communities found at lower resolutions.

2.2. The Gene Ontology

Ontologies are utilized across many disciplines including economics, artificial intelligence, engineering, library science, and biomedical informatics (for a few examples, see [5, 19, 27]). They are a structural framework for organizing information, representing knowledge and describing the relationship between different ideas. Further, they provide a method for conceptualizing common features shared across multiple cases. The Gene Ontology, specifically, describes the relationships between different biological concepts or functions [4]. It breaks these concepts into three main domains, or distinct ontologies: “Biological Process” (BP), describing sets of molecular events, “Molecular Function” (MF), describing the activities of gene products, and “Cellular Component” (CC), describing parts of a cell or its external environment. Each of the three primary domains in GO takes the form of a directed acyclic graph (DAG), in which “child” functional categories, or “terms”, are subclassified under one or more “parent” terms, using “is a” and “part of” relationships. Each parent and all its subsequent progeny therefore define multiple, overlapping, sets of terms, or “branches” in GO. Note that a child term classified in one of the primary domains cannot have a parent classified in a different primary domain. Further the primary domains vary greatly in size with the BP domain containing approximately two-thirds of all terms. Using GO, genes are annotated to terms representing their particular role in a cell, and these annotations are transitive up the relationships in the DAG such that each “parent” term takes on all the gene annotations associated with any of its progeny [30].

In the following analysis we explore if there exists another, natural way to classify terms independent from the ontology structure. To begin, we use term-term ontology relationships and gene-term annotation information for human genes downloaded from the GO website (geneontology.org). We determine the terms with at least one human gene annotation, and plot the cumulative distribution of the size of the branches corresponding to each of these terms and its annotated-progeny (Figure 2(a)). The heavy-tailed distribution is a result of the hierarchical DAG structure as the members of each branch are also members of the parent branch(es). Be-

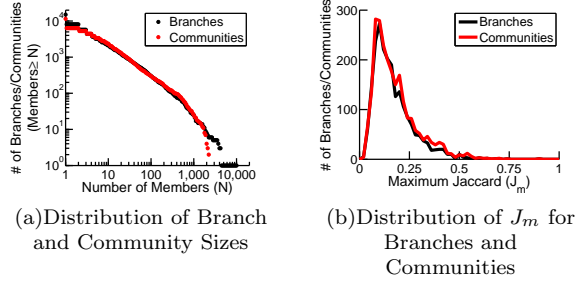


FIG. 2: A comparison of branches in the GO DAG and term communities found by partitioning the term network. (A) The cumulative Distribution for the sizes of all branches in the Gene Ontology and all unique term communities found at the various resolutions. These distributions demonstrate that the number and sizes of branches and communities is similar. (B) Distribution of J_m , the maximum similarity a community or branch with ten or more members has compared to all other branches or communities with ten or more members, respectively. Although a small number of communities and branches have similar memberships, most are highly dissimilar.

cause of the transitive nature of gene annotations, larger branches, which generally represent broader terms, often have many genes annotated to them, whereas terms with fewer or no progeny generally represent more specific biological activities that are performed by relatively fewer genes. Therefore, as with the branch size, a distribution of the number of genes annotated to each term follows a heavy-tailed relationship; however, the number of terms to which each gene is annotated does not appear heavy-tailed (see [11, 12]).

Next, we used the annotation information to construct a gene-term bipartite network. We represent this network in the form of an $n_G \times n_T$ adjacency matrix, where n_G is the total number of genes and n_T is the total number of terms listed in the annotation file. In this matrix a value of one indicates a known connection between the corresponding gene and term, and a value of zero indicates that the gene is not associated with that term. Thus,

$$B_{pi} = \begin{cases} 1 & \text{if gene } p \text{ is annotated to term } i \\ 0 & \text{if gene } p \text{ is not annotated to term } i \end{cases} \quad (2)$$

This bipartite network represents a human-specific summary of the relationships between genes and terms.

We note that although GO is broken into three primary domains and gene-annotations are made to the ontology for many species, for simplicity in the following analysis we will combine information from all three domains and use annotation information only that pertains to human genes.

3. METHODS

3.1. Projecting Term Networks based on Gene Ontology annotations

In this section we use gene-term annotations to construct a network representing term-term relationships. Using the bipartite network (Equation 2) one might create a term network by simply joining together any pair of terms that share common genes; however, this approach would lose a large amount of information as connections between high degree terms would be given the same weight as connections between low degree terms. We correct for the skewed term degree distribution by constructing a diagonal weighting matrix, with elements:

$$w_{ij} = \frac{\delta_{ij}}{\sum_{q=1}^{n_G} B_{qi}} \quad (3)$$

or simply one over the degree of the term in the bipartite graph. Using w , we can project a term-network, T , whose edges are modified by this weighting matrix:

$$T = w' B' B w \quad T_{ij} = \frac{\sum_q B_{qi} B_{qj}}{\sum_q B_{qi} \sum_q B_{qj}} \quad (4)$$

In this network the weights of T have a maximum value of one when the same single gene is annotated to both term i and term j and a minimum value of zero when none of the genes annotated to term i are annotated to term j . The use of the weighting matrix biases the weights of network edges to those between low degree terms and therefore the lower branches of the DAG.

3.2. Identifying Communities of GO terms

Our first goal is to identify the community structure in annotation-driven term-term relationships, meaning we wish to find clusters of terms within which there are many or high-weight relationships in our projected network, but between which there are only few or low-weight relationships. To determine the community structure of the network we used a weighted version of the Fast Greedy Community Structure algorithm [6] and determined the communities of terms at maximum modularity. Fifty-six communities were found with a modularity value of $Q = 0.57$, indicating that there are indeed many individual modules of terms which share gene annotations.

The branches in GO represent multiple, overlapping sets of functional categories rather than one discreet partition of terms, indicative of functional structure at many different levels of specificity. Since standard community structure approaches only find one partition of network

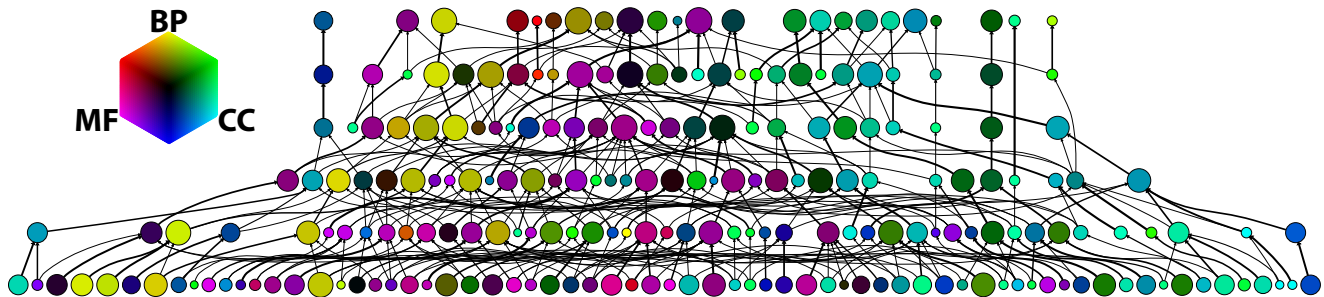


FIG. 3: Visualization of communities (circles) of GO terms found at the six lowest levels of resolution (rows), in increasing order (top to bottom). Width of line connecting two communities is proportional to the percentage of terms in the child community that are also in the parent community. The size of communities is proportional to the log of the number of terms in the community. Color represents the normalized percentage of terms in the community which belong to the BP (yellow), MF (magenta) and CC (cyan) primary domains.

nodes (in our case terms), we implemented a modified version of the Fast Greedy that maximizes modularity for non-zero values of the resolution parameter (see Equation 1) in order to find many different viable partitions. We varied this parameter several orders of magnitude to find 11491 different communities. The different values of the resolution parameter were chosen to give community sizes that were roughly similar to those defined by the branches at different levels of the GO DAG. Like GO branches at different levels, term communities at different resolutions are highly overlapping. We gave our communities numeric identities that vary from TC:0000001 to TC:0011491 and will refer to them as such in the following analysis. The cumulative distribution of the number of members in these communities is shown in Figure 2(a). As with the GO branches, this is a heavy-tailed distribution.

4. RESULTS: AN ALTERNATE “NATURAL” GROUPING OF GO TERMS

4.1. Illustrating Community Structure at different resolutions

To better understand the relationships between the communities found at different resolutions, we visualized the term communities with ten or more members for the six lowest values of resolution used (Figure 3). In this visualization each community is represented by a single circle, whose radius scales as the log of the number of terms belonging to that community and whose color corresponds to the percentage of members from each primary domain that belong to that community. Between the communities found at adjacent resolutions, we draw a line from a community at a higher resolution to a community at a lower resolution if at least 10% of the members of the community from the higher resolution also belong to the community at the lower resolution. The thickness of the line is indicative of the overlap between the two

communities.

While partitions at low resolutions are somewhat similar, we find that high resolution partitions are vastly different. Communities at higher resolutions do not merely represent the “splitting apart” of communities at lower resolutions (represented by a child community only connecting to a single parent), but instead each resolution often brings about a new way of partitioning the network. Sometimes members from multiple communities found at lower resolutions split and/or combine with other lower-resolution communities to form a new community at a higher resolution.

The colors of the communities illustrate how the structure of annotation-driven term relationships is distinct from the structure of those relationships as defined by GO branches. Each GO branch can only belong to one primary ontology, and thus would be pure cyan, magenta or yellow in this type of visualization, however, communities, even smaller ones and those found at higher resolutions, generally contain members from multiple ontologies, resulting in a rainbow of colors.

4.2. Comparing Term Communities and GO Branches

Next we compared the term communities with branches of the GO DAG. In order to quantify the specific differences between the communities and branches, we directly compared the membership of each community with the membership of all branches in GO using the Jaccard similarity (J), calculating a similarity value for every community-branch pair that takes the value:

$$J(x, y) = \frac{|x \cap y|}{|x \cup y|}. \quad (5)$$

For each community (x), we determined the corresponding branch (y) that has the highest overlap in member-

ship by this measure:

$$J_m(x) = \max_{y \in Y} J(x, y), \quad (6)$$

and vice versus.

Because the exact value of the Jaccard similarity is highly sensitive to incremental changes in set membership when comparing sets with only a few members, we will limit all the following analysis to communities and branches that contain ten or more terms in order to focus on the most robust results. Figure 2(b) shows the distribution of J_m comparing these 2370 communities and 2151 branches. Although a handful of communities and branches are quite similar to each other, the majority of communities are dissimilar to the GO Branches and vice versus. To better interpret these values, we selected several communities to inspect more closely.

First we selected a community with a very high J_m value to inspect (Figure 4(a)). TC:0007391 is most similar to GO:0070570 (“regulation of neuron projection regeneration”) with $J_m = 0.6667$. It is interesting that in addition to members from the BP domain TC:0007391 also includes two members from the MF and CC domains, “neutrophin receptor activity” and “perineuronal net” respectively, the former of which is involved in the regeneration of injured axons [10] while the degradation of the latter has been shown to favor axon regeneration [26]. This indicates that terms found in the community but not the branch are consistent with known biology.

Next we selected TC:0001295 and GO:0052547 to illustrate the shared information typically found between a community and the branch (Figure 4(b)). The branch defined by GO:0052547 has members that belong to seven distinct communities, demonstrating that not only are communities often distinct from branches, within the branches themselves the annotation-driven classification is often very distinct from the defined ontological relationships. One might suppose that a significant factor to the dissimilarity found between TC:0001295 and the branches in GO is partially attributable to the fact that it has term members that belong to all of the primary domains. To address this issue for our last example we illustrate a term community (TC:0000936) whose members all belong to the same primary domain (Figure 4(c)). As with TC:0001295 this community only shares a handful of members with its corresponding branch (GO:0060538).

From this analysis we conclude that although there is occasional similarity between our found communities and GO branches, the communities are not simply a recapitulation of the DAG. We have repeated portions of this analysis constructing the term network and corresponding partitions three more times, using only annotations specific to only one of the three primary ontologies, and observe similar results (data not shown).

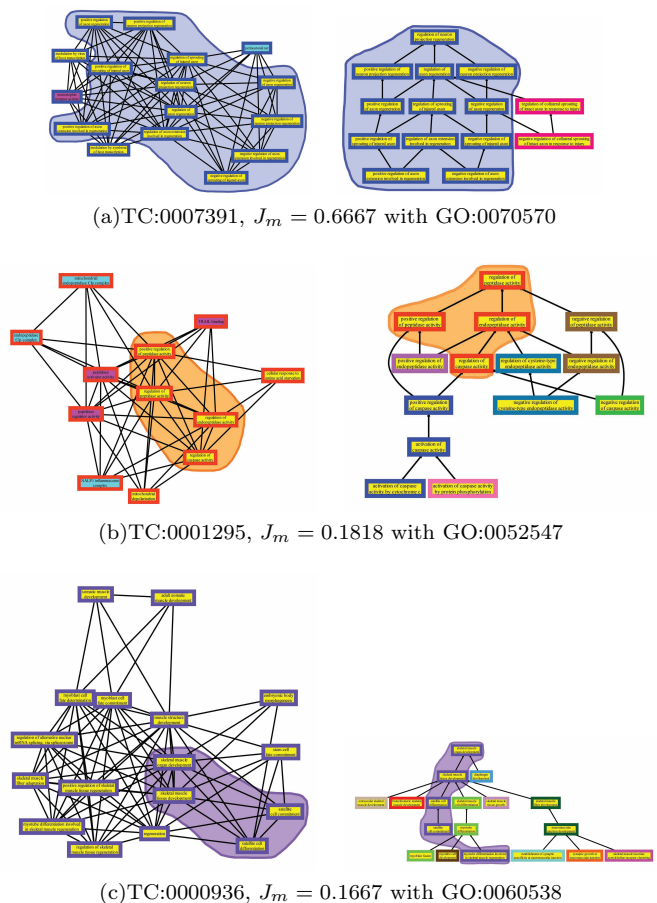


FIG. 4: Three example comparisons between communities and branches: (A) TC:0007391 compared to GO:0070570, (B) TC:0001295 compared to GO:0052547, (C) TC:0000936 compared to GO:0060538. In each panel on the left hand side a community and its inter-community connections in the annotation-driven term network is shown and on the right hand side the branch with which that community has the the highest Jaccard similarity is illustrated. In the right panel edges represent the ontological associations defined by the Gene Ontology term hierarchy. Each term member of the community or branch is colored both by its associated primary domain (inner color - BP:yellow, MF:magenta, CC:cyan) and its community membership (outer color), determined at the same resolution value as the illustrated community. Terms common between the community and the branch are circled.

4.3. Capturing the Biological Information in Term Communities

We have illustrated that our term communities represent a natural partitioning of biological functions that is distinct from the GO DAG, however, the biological meaning of these communities is, at this point, unclear. On a mathematical level they represent sets of biological functions that are generally performed by the same collection of genes. Labeling and understanding the biological meaning behind these communities is vital if they

are to have the same of wide-range applications as the GO branches.

At first one might consider determining the information contained within each community by manually inspecting the functional categories belonging to those communities. For communities containing only a handful of terms, such as those illustrated in Figure 4, this is doable. Unfortunately this task can be daunting at best for larger communities, each of which contains several hundred members. Therefore, in order to easily interpret the contents of our communities we summarize the information contained in each in the form of word clouds (Figure 5(a)). Specifically, for each community, we determine the descriptions corresponding to the member terms of that community. We then count the number of times an individual word appears across all these descriptions and calculate the statistical enrichment of the frequency of that word in the community compared to its frequency across the descriptions of all GO terms. Finally, we use a publicly available word-cloud making program [1] to illustrate the words belonging to each community, scaling the relative sizes of the words based on the word’s statistical enrichment in the community and coloring each word based on the normalized percentage of times the term is derived from is a member of each primary domain.

As an example, take Community TC:0000400. Looking at all 335 members of this community and trying to manually determine what biological information they represent could be difficult or impossible. The word cloud presentation (Figure 5(b)), however, easily summarizes this information and reveals that this community includes biological concepts related to various types of RNA, including “rRNA”, “tRNA”, “mRNA”, “LSU-rRNA”, “SSU-rRNA”, “ncRNA”, “RNA-polymerase” and more. It has members belonging to all three primary domains, and often the individual words shared between those members are associated with multiple domains, resulting in a complex coloration. We note that this community is not highly similar to any particular branch in GO. It has the highest similarity to GO:0016070 or “RNA metabolic process”, with a value of $J_m = 0.22146$. Other word clouds show a similar richness of information. For example TC:0000061 contains many words related to the heart such as “cardiac”, “muscle”, “ventricle”, “ventricular” and “heart.” (Figure 5(c)). This community is also distinct from the branches with a J_m value of 0.149 with GO:0072358.

One can also represent the biological information contained in branches in the form of word clouds, although, because the members of each branch can only belong to one of the three primary domains, all the words in the cloud will be the same color. Two branches (GO:0000003 and GO:0002376) are illustrated in Figure 5(d)-(e). The first clearly contains terms pertaining to sex-related processes as it contains words such as “female”, “sex”, “prostate” and “male.” This corresponds well to the name of the parent term in the branch, “reproduction”. Similarly, the cloud for GO:0002376, whose parent term

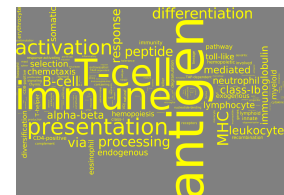
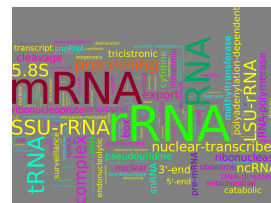


FIG. 5: (A) A schematic illustrating how the biological information contained in a particular community or branch can be summarized in a word cloud. (B-E) Term Communities (TC:0000400, TC:0000061) and branches (GO:0009607, GO:0050896) summarized as word clouds. In each case the color of a word represents how often the term description containing that word belongs to each of the primary domains (BP: yellow, MF: magenta, CC: cyan, also see Figure 2.2 for mixed-domain coloration) and size represents that word’s statistical enrichment in that community/branch.

name is “immune system process” contains words pertaining to the immune system.

4.4. Using the term communities to evaluate and predict genetic function

Finally, we wanted to test how our communities might be used in one common application of the Gene Ontology: functional enrichment analysis. To perform this analysis we chose to use the structure-preserving form of Annotation Enrichment Analysis (SP-AEA) [11] since it has been shown to better estimate the biological functions of experimentally derived sets of genes, and has a conceptual framework conducive to estimating functional enrichment between sets of terms and genes, rather than simply between two sets of genes.

To begin, we downloaded a collection of experimentally derived genes sets from the Gene Signatures Database (GeneSigDB) [8]. This database is a manual curation of previously published gene expression signatures, focusing

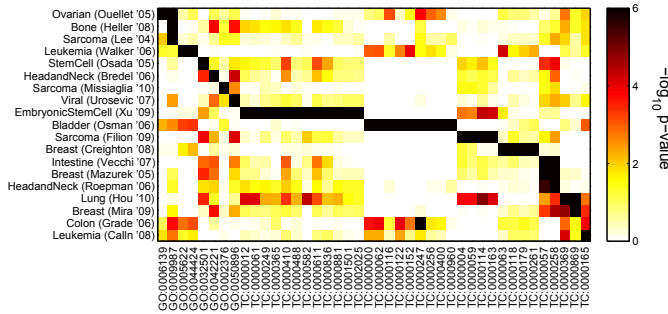


FIG. 6: A heat map showing the statistical enrichment of selected cancer signatures (see text) in GO branches and term communities.

primarily on cancer and stem cell signatures [7]. In the following analysis we will use all 509 human signatures from this database that contain at least 100 and less than 1000 genes annotated in the Gene Ontology. Using SP-AEA (with one million randomizations), we estimated the significance of enrichment for these 509 signatures in both the term communities and GO branches.

Term communities show a level of statistical enrichment equal or greater to that of GO branches, lending biological validity to our term communities. Furthermore this level of enrichment is not evident using random communities (data not shown). We wished to know if there was a context in which our term communities captured biological information missed by the branches, or vice versus. Thus we selected gene signatures that were significantly enriched ($p < 10^{-6}$) in at least one community/branch but not significantly enriched ($p > 5 \times 10^{-5}$) in any GO branch/community. Figure 6 shows a heat map of the nineteen signatures that met this criteria with the communities and branches for which they are enriched.

It is immediately striking that of these signatures, the majority are enriched in communities and not GO branches, rather than the other way around. Of course, some of these communities may capture redundant information and only vary by a few members, however, we point out this is equivalent in structure to the GO branches, where the branch of a term will often contain virtually the same membership as that of its immediate progeny and parent (see, for example, the branch defined by GO:0060538, illustrated in Figure 4(c)).

Two signatures, in particular, are enriched in a collection of communities. The first, an embryonic stem cell signature [32], represents genes that are up-regulated in cardiomyocytes compared to non-selected embryoid bodies and hESC. The communities represented in this signature contain several different themes, all consistent with the expected properties of genes selected from stem cells and related to the heart. The corresponding clouds emphasize words such as “cardiac” and “muscle” (TC:0000012), “actin”, “myosin”, and “fil-

ament” (TC:0000249), “morphogenesis” and “development” (TC:0000365), “blood”, “pressure” and “contraction” (TC:0000582), with the other clouds generally containing these words in different combinations (see for example TC:0000061, illustrated in Figure 5(c)).

The second signature is a list of bladder cancer specific genes [25]. Most of the words emphasized by the community clouds are related to cell proliferation. For example, it is enriched in TC:0000400 (illustrated in Figure 5(b)) and the other clouds emphasize words such as “cell-cycle”, “mitotic”, “meiotic”, “checkpoint”, “repair”, “replication”, “recombination”, “telomere”, “spindle”, “complex”, “DNA”, “chromosome”, “histone”, and “methylation”. Although the connection to the bladder is not obvious, the connection to cancer and the high rate of cell proliferation in tumor cells [2] is apparent.

5. DISCUSSION

The network structure of gene annotations using the Gene Ontology has not previously been exploited in a manner that reveals an organization of biological function that is unique from the published hierarchical classification of the GO DAG. By analyzing functional annotation data we were able to construct an alternate, natural, and biologically-relevant way in which to categorize cellular functions. This categorization is structurally and conceptually distinct from the GO DAG and allows for functional relationships between terms that do not share a parent/child relationship. It takes advantage of a large amount of data from a variety of sources and creates a classification scheme that is motivated primarily by the data reported rather than the organization of human conceptions.

The term communities defined in this work represent an integration of information across all three primary domains in GO that, to the authors’ knowledge, has not previously been investigated systematically. Using the simple principle of co-annotation we suggest that in the future biological concepts from other smaller databases could also be analyzed or even combined with these results. Additionally, we concede that the communities defined here likely do not represent the only way to group functional terms outside of the ontology structure. Even so, we believe that the functional enrichment analysis presented here demonstrates that the term communities we define are clearly more than a mathematical phenomenon and have a high potential to be used to better interpret biological data.

Acknowledgments

We wish to thank Geet Duggal for supplying an implementation of the Fast Greedy Community Structure algorithm that included the resolution parameter.

- [1] Ibm word cloud generation software. URL <http://www.softpedia.com/progDownload/IBM-Word-Cloud-Generator-Download-160283.html>.
- [2] M. Andreeff, G. DW, and P. AB. *Holland-Frei Cancer Medicine*. BC Decker, Hamilton, ON, 5th edition, 2000.
- [3] A. Arenas, A. Fernandez, and S. Gomez. Analysis of the structure of complex networks at different resolution levels. Jan. 2008. doi: 10.1088/1367-2630/10/5/053039. URL <http://dx.doi.org/10.1088/1367-2630/10/5/053039>.
- [4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. 25(1):25-9+, 2000. ISSN 1061-4036. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10802651.
- [5] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 14(1):20-26, Jan. 1999. ISSN 1541-1672. doi: 10.1109/5254.747902. URL <http://dx.doi.org/10.1109/5254.747902>.
- [6] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111+, Dec. 2004. doi: 10.1103/PhysRevE.70.066111. URL <http://dx.doi.org/10.1103/PhysRevE.70.066111>.
- [7] A. C. Culhane, T. Schwarzl, R. Sultana, K. C. Picard, S. C. Picard, T. H. Lu, K. R. Franklin, S. J. French, G. Papenhausen, M. Correll, and J. Quackenbush. GeneSigDB—a curated database of gene expression signatures. *Nucleic acids research*, 38(Database issue):D716-D725, Jan. 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp1015. URL <http://dx.doi.org/10.1093/nar/gkp1015>.
- [8] A. C. Culhane, M. S. Schröder, R. Sultana, S. C. Picard, E. N. Martinelli, C. Kelly, B. Haibe-Kains, M. Kapushesky, A.-A. St Pierre, W. Flahive, K. C. Picard, D. Gusenleitner, G. Papenhausen, N. O'Connor, M. Correll, and J. Quackenbush. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research*, 40(D1):D1060-D1066, Jan. 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr901. URL <http://dx.doi.org/10.1093/nar/gkr901>.
- [9] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American journal of human genetics*, 78(6):1011-1025, June 2006. ISSN 0002-9297. doi: 10.1086/504300. URL <http://dx.doi.org/10.1086/504300>.
- [10] H. Funakoshi, J. Frisén, G. Barbany, T. Timmusk, O. Zachrisson, V. M. Verge, and H. Persson. Differential expression of mRNAs for neurotrophins and their receptors after axotomy of the sciatic nerve. *The Journal of cell biology*, 123(2):455-465, Oct. 1993. ISSN 0021-9525. URL <http://view.ncbi.nlm.nih.gov/pubmed/8408225>.
- [11] K. Glass and M. Girvan. Annotation enrichment analysis: An alternative method for evaluating the functional properties of gene sets, Aug. 2012. URL <http://arxiv.org/abs/1208.4127>.
- [12] K. Glass, E. Ott, W. Losert, and M. Girvan. Implications of functional similarity for gene regulatory interactions. *Journal of the Royal Society, Interface / the Royal Society*, Feb. 2012. ISSN 1742-5662. doi: 10.1098/rsif.2011.0585. URL <http://dx.doi.org/10.1098/rsif.2011.0585>.
- [13] R. Guimera and L. A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895-900, Feb. 2005. ISSN 0028-0836. doi: 10.1038/nature03288. URL <http://dx.doi.org/10.1038/nature03288>.
- [14] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucl. Acids Res.*, 35(Web Server issue):gkm415+, June 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm415. URL <http://dx.doi.org/10.1093/nar/gkm415>.
- [15] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41-42, May 2001. ISSN 0028-0836. doi: 10.1038/35075138. URL <http://dx.doi.org/10.1038/35075138>.
- [16] O. D. King, R. E. Foulger, S. S. Dwight, J. V. White, and F. P. Roth. Predicting gene function from patterns of annotation. *Genome research*, 13(5):896-904, May 2003. ISSN 1088-9051. doi: 10.1101/gr.440803. URL <http://dx.doi.org/10.1101/gr.440803>.
- [17] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555-1558, Nov. 2004. ISSN 1095-9203. doi: 10.1126/science.1099511. URL <http://dx.doi.org/10.1126/science.1099511>.
- [18] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Physical Review Letters*, 100:118703+, Mar. 2008. doi: 10.1103/PhysRevLett.100.118703. URL <http://dx.doi.org/10.1103/PhysRevLett.100.118703>.
- [19] U. Mäki, editor. *The Economic World View: Studies in the Ontology of Economics*. The Press Syndicate of the University of Cambridge, Cambridge, UK, 2001.
- [20] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824-827, Oct. 2002. ISSN 1095-9203. doi: 10.1126/science.298.5594.824. URL <http://dx.doi.org/10.1126/science.298.5594.824>.
- [21] S. Mostafavi and Q. Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, 26(14):1759-1765, July 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq262. URL <http://dx.doi.org/10.1093/bioinformatics/btq262>.
- [22] M. E. Newman. Analysis of weighted networks. 70 (5 Pt 2):056131+, 2004. ISSN 1539-3755. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15600716.

- [23] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. 69(2 Pt 2):026113+, 2004. ISSN 1539-3755. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14995526.
- [24] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. URL <http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=SIREAD0000450000020001670000001&idtype=cvips&gifs=yes>.
- [25] I. Osman, D. F. Bajorin, T.-T. T. Sun, H. Zhong, D. Douglas, J. Scattergood, R. Zheng, M. Han, K. W. Marshall, and C.-C. C. Liew. Novel blood biomarkers of human urinary bladder cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 12(11 Pt 1):3374–3380, June 2006. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-05-2081. URL <http://dx.doi.org/10.1158/1078-0432.CCR-05-2081>.
- [26] E. Pastrana, M. T. T. Moreno-Flores, E. N. Gurzov, J. Avila, F. Wandosell, and J. Diaz-Nido. Genes associated with adult axon regeneration promoted by olfactory ensheathing cells: a new role for matrix metalloproteinase 2. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(20):5347–5359, May 2006. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.1111-06.2006. URL <http://dx.doi.org/10.1523/JNEUROSCI.1111-06.2006>.
- [27] S. B. Shum, E. Motta, and J. Domingue. ScholOnto: an Ontology-Based digital library server for research documents and discourse. *International Journal on Digital Libraries*, 3:237–248, 2000. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.3835>.
- [28] R. V. Solé, R. F. Cancho, J. M. Montoya, and S. Valverde. Selection, tinkering, and emergence in complex networks. *Complex.*, 8(1):20–33, Sept. 2002. ISSN 1076-2787. URL <http://portal.acm.org/citation.cfm?id=770715.770720>.
- [29] R. Stevens, C. A. Goble, and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, 1(4):398–414, January 2000. doi: 10.1093/bib/1.4.398. URL <http://dx.doi.org/10.1093/bib/1.4.398>.
- [30] The_gene_ontology_consortium. Creating the gene ontology resource: design and implementation. *Genome Res.*, 11(8):1425–1433, Aug. 2001. ISSN 1088-9051. doi: 10.1101/gr.180801. URL <http://dx.doi.org/10.1101/gr.180801>.
- [31] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18(7):1283–1292, July 2001. ISSN 0737-4038. URL <http://mbe.oxfordjournals.org/cgi/content/abstract/18/7/1283>.
- [32] X. Q. Q. Xu, S. Y. Y. Soo, W. Sun, and R. Zweigerdt. Global expression profile of highly enriched cardiomyocytes derived from human embryonic stem cells. *Stem cells (Dayton, Ohio)*, 27(9):2163–2174, Sept. 2009. ISSN 1549-4918. doi: 10.1002/stem.166. URL <http://dx.doi.org/10.1002/stem.166>.
- [33] X. Yang, Y. Zhou, R. Jin, and C. Chan. Reconstruct modular phenotype-specific gene networks by knowledge-driven matrix factorization. *Bioinformatics*, 25(17):2236–2243, Sept. 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp376. URL <http://dx.doi.org/10.1093/bioinformatics/btp376>.
- [34] A. Youn, D. J. Reiss, and W. Stuetzle. Learning transcriptional networks from the integration of ChIP-chip and expression data in a non-parametric model. *Bioinformatics*, 26(15):1879–1886, Aug. 2010. doi: 10.1093/bioinformatics/btq289. URL <http://dx.doi.org/10.1093/bioinformatics/btq289>.
- [35] J. Zhao, H. Yu, J. Luo, Z. Cao, and Y. Li. Complex networks theory for analyzing metabolic networks. *Chinese Science Bulletin*, 51(13):1529–1537–1537, July 2006. ISSN 1001-6538. doi: 10.1007/s11434-006-2015-2. URL <http://dx.doi.org/10.1007/s11434-006-2015-2>.